



Understanding the Query: THCIB and THUIS at NTCIR-10 Intent Task

Yunqing Xia¹ and Sen Na²

¹ Tsinghua University

² Canon Information Technology (Beijing) Co. Ltd.

Before we start

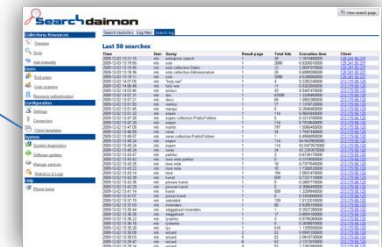
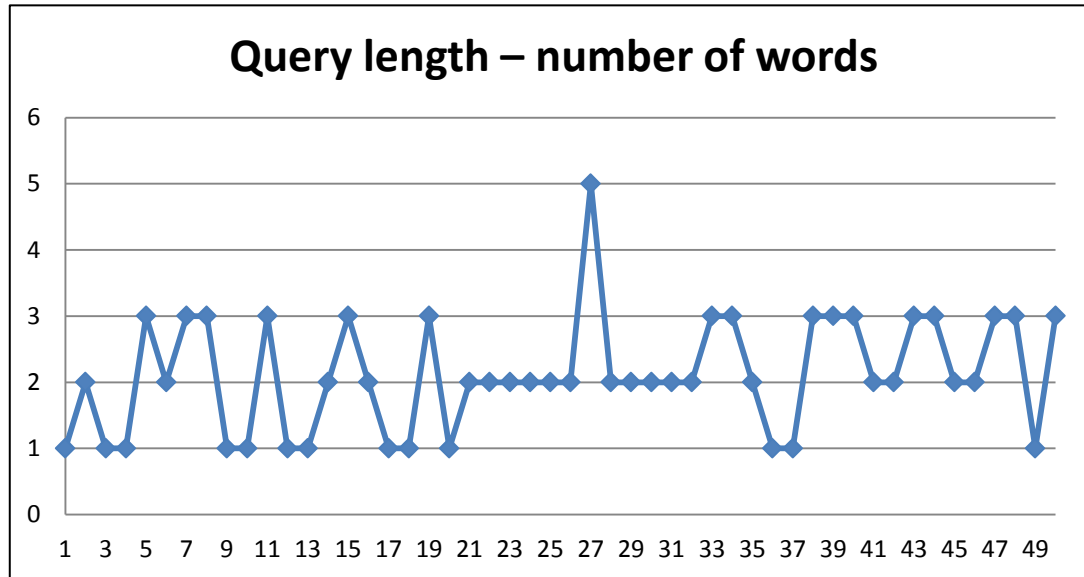
- Who are we?
 - **THUIS** is the research team at Intelligent Search group at Center for Speech and Language Technology, Tsinghua University
 - **THCIB** is the joint research team between THUIS and Canon Information Technology (Beijing) Co. Ltd..
- Why did we participate NTCIR INTENT task?
 - We believe intent mining is one of the most promising technologies to make the search engines smarter thus more helpful to human.
 - We view query-based intent mining as a major topic in our research group
- What is task/subtask we participated?
 - Subtopic mining: Systems are required to return a ranked list of *subtopic strings* in response to a given topic query while the top N subtopic strings should be *both relevant and diversified* as much as possible.

Outline

- The motivation
- System overview
- What make our system different?
- Evaluation
 - The submitted runs
 - Results and discussion
- Conclusion and future work

The Motivation (1/3)

- ISSUE #1: Query is usually very short



- SOLUTION #1: Applying BIGGER CONTEXT in query understanding
 - General knowledge base: Wikipedia
 - User behavior data: Query log, search engine auto-completions and suggestions
 - Search results: Title and snippet

The Motivation (2/3)


- **ISSUE #2: Subtopic surface strings are redundant**

furniture for small spaces **store**
{furniture for small spaces **market**
{furniture for small spaces **wholesale**
{furniture for small spaces **shop**
{furniture for small spaces **center**


.....

{furniture for small spaces **Tokyo**
furniture for small spaces **New York**
{furniture for small spaces **London**
{furniture for small spaces **Hong Kong**
{furniture for small spaces **Indonesia**

.....



furniture for small spaces
store {**store, market,**
wholesale, shop, center, ...}

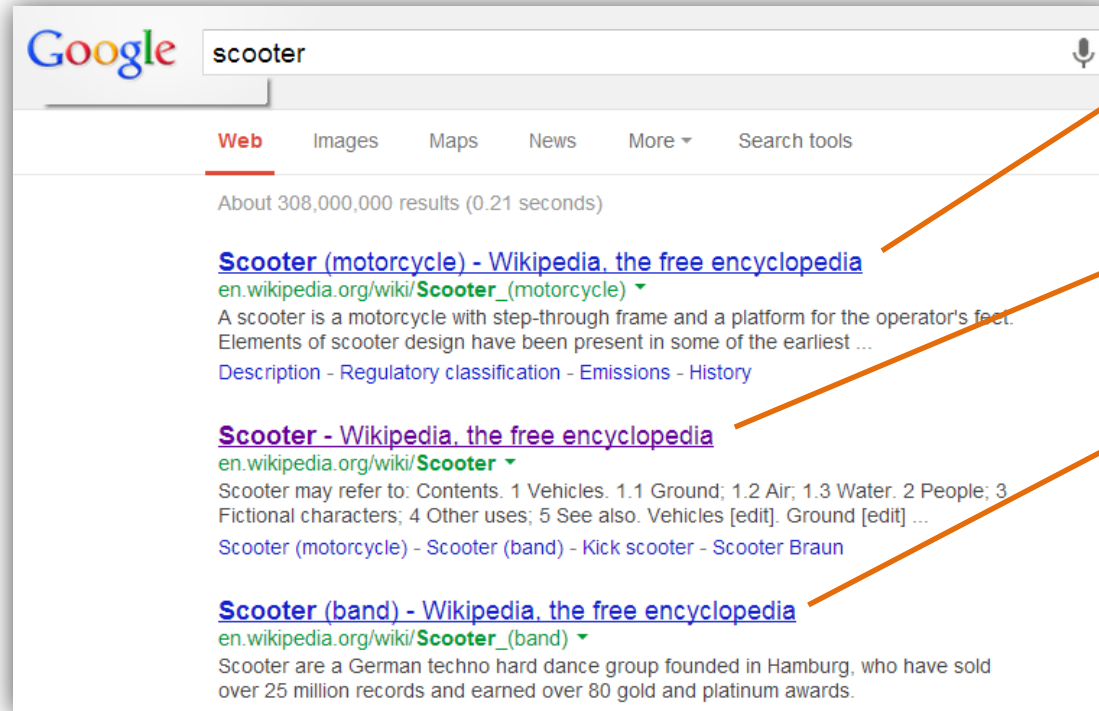


furniture for small spaces
Tokyo {**Tokyo, New York,**
London, Hong Kong,
Indonesia, ...}

- **SOLUTION #2: Discover the implicit intents by clustering the subtopic surface strings**
 - A sense-based clustering algorithm

The Motivation (2/3)

- ISSUE #3: Relevance is no longer effective for intent ranking



Wikipedia

Wikipedia again!!

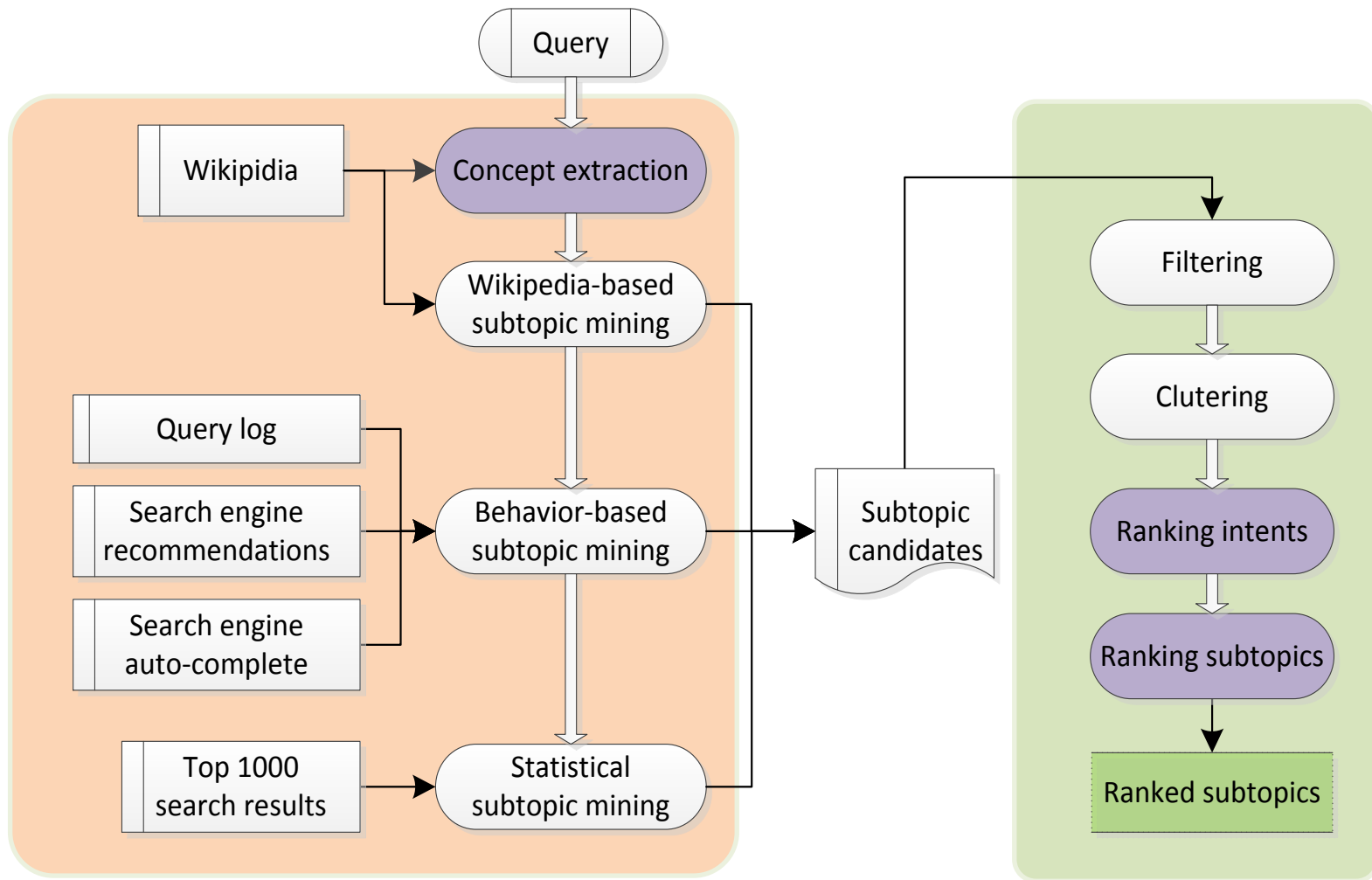
Still Wikipedia!!

- SOLUTION #3: Ranking intents considering both relevance and diversity
 - A unified intent weighting model and a subtopic selecting strategy

Outline

- The motivation
- **System overview**
- What make our system different?
- Evaluation
 - The submitted runs
 - Results and discussion
- Conclusion

System overview



Subtopic candidate mining (SCM)

Subtopic candidate ranking (SCR)

Outline

- About the NTCIR10 INTENT-2 task
 - Who are we?
 - Why do we participate NTCIR INTENT task?
 - Task/subtask we participated
- The motivation
- System overview
- **What make our system different?**
- Evaluation
 - The submitted runs
 - Results and discussion
- Conclusion and future work

What make our system different?

- Concept based
 - Wikipedia entries and related entries
 - From query analysis to expansion
 - From subtopic extraction to intent mining
 - From relevance to diversity
 - From weighting to ranking
- Discovering intent for diversification
 - Word sense induction
 - Intent induction/disambiguation
 - Entity analysis to address homogeneous exclusive subtopics

SCM: Extracting concepts from query

- Downloading the entire Wikipedia
 - Entry ==> Concept
 - Concept ==> Definition
 - Concept → Related concepts
- Bi-directional maximum entry matching
- Using the multiple matches in the disambiguation page
- Using redirects when no entry is exactly matched



“battles in the civil war” → “battle”, “civil war”

Battle
From Wikipedia, the free encyclopedia
(Redirected from Battles)

This article is about combat. For other uses, see *Battle* (disambiguation).

Generally, a **battle** is a conceptual component in the hierarchy of combat in warfare between two or more armed forces, or combatants. A war sometimes consists of many battles. Battles generally are well defined in duration, area and force commitment.^[1] Wars and military campaigns are guided by strategy, whereas battles take place on a level of planning and execution known as operational mobility.^[2] German strategist Carl von Clausewitz stated that "the employment of battles ... to achieve the object of war^[3] was the essence of strategy."

Contents [hide]

- Etymology
- Characteristics
- Battlespace
- Factors
- Types
 - Land
 - Naval
 - Aerial
- Timing
- Effects
- See also
- References

Etymology [edit]

The definition of a battle cannot be arrived at solely through the names of historical battles, many of which are misnomers. The word *battle* is a loanword in English from the Old French *bataille*, first attested in 1207, from Late Latin *battalia*, meaning "exercise of soldiers and gladiators in fighting and fencing", from Late Latin (taken from Germanic) *battwære* "beat", from which the English word *battery* is also derived.



Civil war
From Wikipedia, the free encyclopedia

This article is about the definition of the specific type of war. For civil wars in history, see *List of civil wars*. For other uses, see *Civil war* (disambiguation).

A **civil war** is a war between organized groups within the same nation state or empire,^[1] or, less commonly, between two countries created from a formerly united nation state.^[2] The aim of one side may be to take control of the country as a nation, to achieve independence for a region, or to change government policies.^[3] The term is a calque of the Latin *bellum civile*, which was used to refer to the various civil wars of the Roman Republic in the 1st century BC.

A civil war is a high-intensity conflict, often involving regular armed forces, that is sustained, organized and hierarchical. Civil war may result in large numbers of casualties and the consumption of significant resources.^[3]


Civil wars since the end of World War II have lasted on average just over four years, a dramatic rise from the one-and-a-half year average of the 1900-1964 period. While the rate of occurrence of new civil wars has been relatively steady since the mid-19th century, the increasing length of those wars resulted in increasing numbers of war deaths per year. For example, there were no more than five civil wars underway simultaneously in the first half of the 20th century, while over 20 concurrent civil wars were occurring at the end of the Cold War, before a significant decrease in conflicts strongly associated with the superpower rivalry came to an end. Since 1945, civil wars have resulted in the deaths of over 25 million people, as well as the forced displacement of millions more. Civil wars have further resulted in economic collapse. Burma (Myanmar), Uganda and Angola are examples of nations that were considered to have promising futures before being engulfed in civil wars.^[4]

Contents [hide]

- Formal classification
- Causes of civil war in the Collier-Rodrik Model
- Other causes
- Duration of civil wars
 - Civil wars in the 19th and early 20th centuries
 - Civil wars since 1945
 - Effect of the Cold War
- See also
- References
- Bibliography
- External links

Formal classification

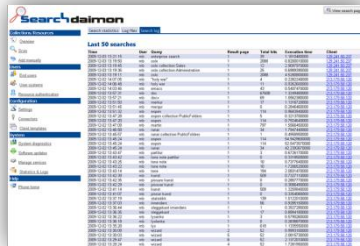
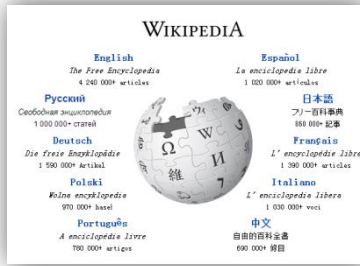
James Fearon, a scholar of civil wars at Stanford University, defines a civil war as "a violent conflict within a country fought by organized groups that aim to take power at the center or in a region, or to change government policies."^[1] Jon Brunschweiler further specifies that one side of a civil



SCM: Query expansion

- Wikipedia
 - Synonymous entries (redirects) and the related concepts
 - Polysemous entries (disambiguation pages)
- Intent schema
 - {concepts, prepositions, wild cards}
 - “hobby store”: “* of hobby store”, “* at hobby store”, “hobby store in *”, “hobby store at *”, etc.
- Concept repositioning
 - “battles in civil war” → “battles civil war”, “civil war battles”
- The motivation:
 - Reforming the query so as to obtain subtopic candidates as many as possible (in query auto-completion, query suggestions, etc.)

SCM: Extracting subtopic candidates



Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon	Searchdaimon
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9
10	10	10	10	10	10	10	10	10	10



- **Wikipedia** – general knowledge base
 - Concept definition
- **User Behavior Data** – user centric data
 - Co-occurrence
 - Search engine tools (auto-completion, query suggestion)
- **Search Results** – pseudo feedback
 - Query topics (word senses) within snippets of top N=1000 results

Wikipedia concept definition

subtopic

- **Ground**
 - [Kick scooter](#)
 - [Motorized scooter](#)
 - [Scooter \(motorcycle\)](#)
 - [Knee scooter](#)
 - [Mobility scooter](#)
 - [Eccentric-hub scooter](#)
 - [Square scooter](#)
- **Air**
 - [Douglas A-4 Skyhawk](#)
 - [Air Scooter](#)
- **Water**
 - [Underwater scooter](#)
 - [Water scooter](#)
 - [Ice boat](#)
- **People**
 - [Scooter Braun](#)
 - [Lloyd L. Burke](#) (nicknamed "Scooter")
 - [Dill Stokes](#) (nicknamed "Scooter")
- **Fictional characters**
 - [Scooter \(comics\)](#)
 - [Scooter \(Coronation Street\)](#)
 - [Scooter \(Gobots\)](#)
 - [Scooter \(Muppet\)](#)
 - [Scooter \(SpongeBob SquarePants\)](#)
 - [Scooter \(talking baseball\)](#)
 - [Scooter: Secret Agent](#)

Scooter

From Wikipedia, the free encyclopedia
(Redirected from *Scooters*)

Scooter may refer to:

Vehicles [[edit](#)]

Ground [[edit](#)]

- [Kick scooter](#), a vehicle propelled by a standing rider pushing off the ground
- [Motorized scooter](#), a motorized version of a kick scooter
- [Scooter \(motorcycle\)](#), a motorcycle with a step-through frame and a platform for the feet
- [Knee scooter](#), a mobility device used for walking by people with leg injuries
- [Mobility scooter](#), a motorized chair
- [Eccentric-hub scooter](#), a two-wheeled vehicle propelled by a bouncing rider
- [Square scooter](#), a square plank with four swivel casters

Air [[edit](#)]

- [Douglas A-4 Skyhawk](#), a ground-attack aircraft
- [AirScooter](#), a theoretical ultralight helicopter

Water [[edit](#)]

- [Underwater scooter](#), a piece of diving equipment
- [Water scooter](#), a recreational watercraft
- [Ice boat](#), a vehicle for quick travel across water, ice or snow

People [[edit](#)]

- [Scooter Braun](#) (born 1981), American talent manager
- [Lloyd L. Burke](#) (1924–1999), U.S. Army soldier, nicknamed "Scooter"
- [Oll Trigg](#) (born 1950), American political figure
- [Dill Stokes](#) (1917–2007), American baseball player, nicknamed "The Scooter"

Fictional characters [[edit](#)]

- [Scooter \(comics\)](#)
- [Scooter \(Coronation Street\)](#)
- [Scooter \(Gobots\)](#)
- [Scooter \(Muppet\)](#)

User Behavior Data

- The user search log (e.g., ClueWeb09)
- Tools of commercial search engines based on user behavior data
 - Auto-completion
 - Query suggestion
- With expanded queries based on **concepts**

Search results

- Search with **concept** as a whole keyword
 - In query <battles in the civil war>, <“civil war”> is **one** keyword *WORD*
 - In Web pages, <“civil war”> is **one** keyword *WORD* (‘war’ must immediately follow ‘civil’)
- Induce aspects of the query using **WSI** (**word sense induction**) technique
 - LDA + keyword extraction
 - Labeled LDA
 - *Sense based LDA: a sense based clustering algorithm*

Reference: A paper submitted to CIKM 2013.

SCR: Re-calculating the relevance score

- Replacing bag of word with bag of Wikipedia concepts
 - BM25 again.
- Incorporating source score

$$p_t = w_{ST}(t) + w_{SC}(t)$$

- $w_{st}(t)$: Relevance score of the subtopic candidates
- $w_{sc}(t)$: Importance score (empirical) of the source where the subtopic comes from.

SCR: Discovering intents

- Clustering subtopics candidates with Affinity Propagation (AP) algorithm
 - Calculating subtopic similarity with VSM-based cosine similarity
 - Extraction concept-based VSM features from snippets of the top 50 search results with subtopic string as a query.
 - Choosing mean of the similarity matrix as clustering preference value
- The revised version
 - Choosing mean of the subtopic importance value (=relevance + resource weight)

SCR: Weighting the intents

- A simple sum equation

$$w_{IN} = \sum_{i=1}^N [w_{ST}(t_i) + w_{SC}(t_i)]$$


SCR: Entity analysis

- Homogenous exclusive entities are found many in subtopic candidates
 - “furniture for small spaces **New York**”
 - “furniture for small spaces **Los Angeles**”
- Freebase - a global resource of ontology
 - It provides HTTP API for data retrieval
 - The whole dump data can be downloaded from Web
- Judgment of homogenous exclusive entities
 - Sharing the same immediate father node!
 - “**City/Town/Village**”

Freebase Find... Browse Query Help

Search


Search Results

 **New York City** /m/02_286
City/Town/Village, Location, Filming location, Fictional Setting, Statistical region, Dated location
Governmental Jurisdiction, Organization scope
country: United States of America
time zones: Eastern Time Zone
area: 1213.4
population: 8336697 (2012), 8244910 (2011), 8175133 (2010), 8391881 (2009), 8346794 (2008), 8213839 (2005), 8169940 (2004), 8126718 (2003), 8092749 (2002), 8063137 (2001), 8015347382901 (1997), 7360622 (1996), 7349560 (1995), 7341300 (1994), 7329079 (1993), 7304897894862 (1990), 7894862 (1970), 7781984 (1960), 7891957 (1950), 7454995 (1940), 6930443437202 (1900), 1515301 (1890), 1206299 (1880), 942292 (1870), 813669 (1860), 515547 (1850)

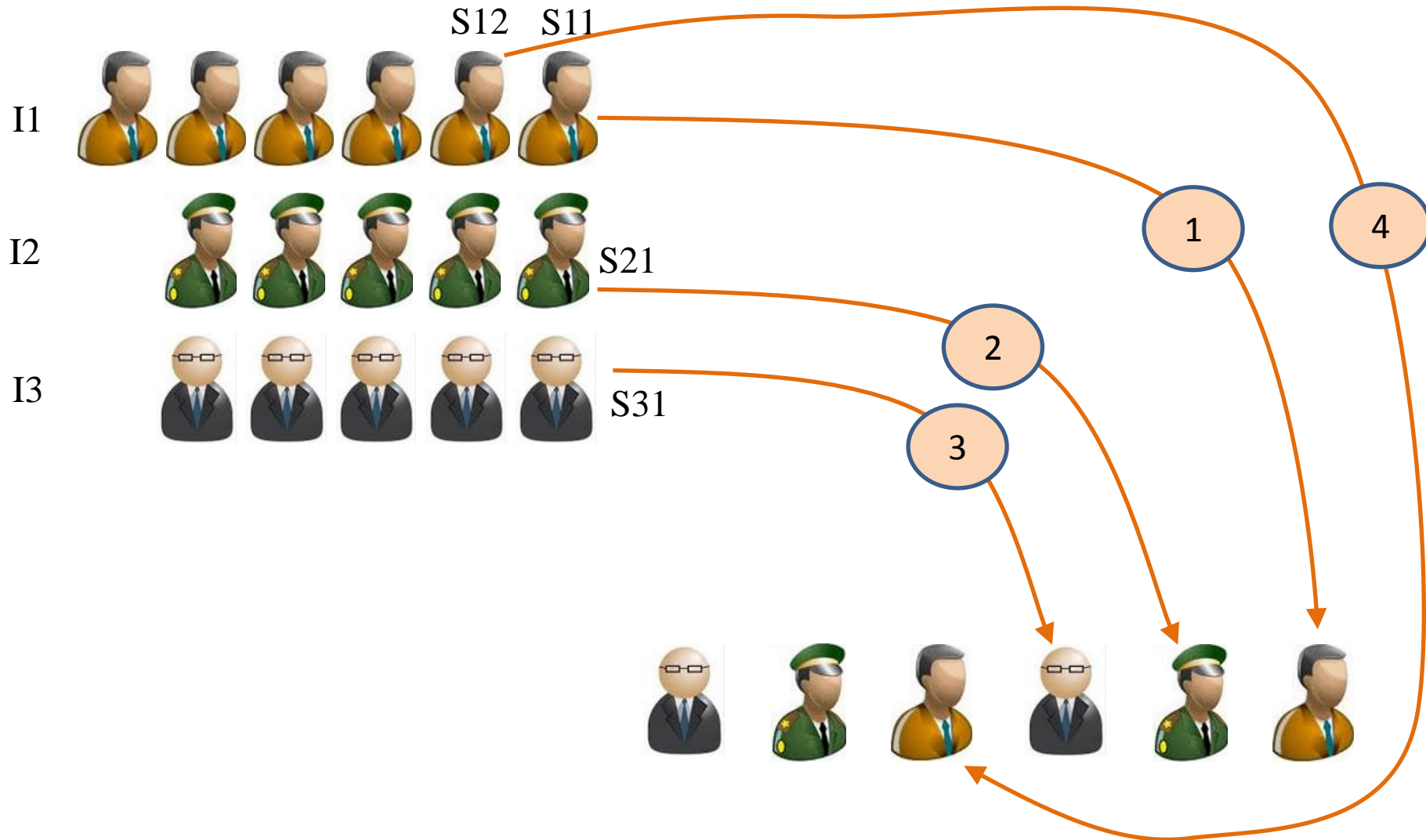
Freebase Find... Browse Query Help

Search

Search Results

 **Los Angeles** /m/030qb3t
City/Town/Village, Olympic host city, Travel destination, Governmental Jurisdiction, HUD County
time zones: Pacific Time Zone
area: 1301.97
population: 3819702 (2011), 3792621 (2010), 3831868 (2009), 3801576 (2008), 3778658 (2007), 3703921 (2000), 3633591 (1999), 3598002 (1998), 3564547 (1997), 3541941 (1996), 353432816061 (1970), 2479015 (1960), 1970358 (1950), 1504277 (1940), 1238048 (1930), 576677 (1920)
alias: Los Angeles, California, L.A., City of Angels, Los Angeles, USA, Los Angeles, California
Los Angeles, officially the City of Los Angeles, often known by its initials L.A., is the most populated city in California and the second most populated city in the United States. It has a population of 3,819,702 as of the 2010 United States Census of 3,792,621. It has an area of 469 square miles (1,213.4 square kilometers).

SCR: Selecting for ranking



Outline

- The motivation
- System overview
- What make our system different?
- **Evaluation**
 - The submitted runs
 - Results and discussion
- Conclusion and future work

The submitted runs

- We submitted 5 runs for English task

RUN ID	Description
<i>THCIB-S-E-1A</i>	SCM (1.Concept extraction + 3.Wikipedia + 4.Query log+ 5.Search results + 6.Filtering) + SCR (similarity + source importance+ relevance)
<i>THCIB-S-E-2A</i>	THCIB-S-E-1A + SCM (2.Query expansion)
<i>THCIB-S-E-3A</i>	THCIB-S-E-2A +SCR (4.Entity analysis)
<i>THCIB-S-E-4A</i>	THCIB-S-E-3A +SCR (3.Intent discovering with standard AP + 5.Intent weighting+ 6. Subtopic selecting)
<i>THCIB-S-E-5A</i>	THCIB-S-E-4A + SCR (3.Intent mining with revised AP)

- We submitted 4 runs for Chinese task
 - No Freebase in Chinese (Run 3 in English is not planned for Chinese task).

Results and discussion – Performance

- Rank:
 - Run 2 > Run 1 > Run 3 > Run 5 > Run 4
- Observations
 - Concept-based query expansion is useful in subtopic mining (Run 2 vs. 1)
 - Entity analysis is not appropriately used (Run 4)
 - Performance of intent discovery can be improved (Run 3)
 - Intent-based subtopic weighting model can be improved (Run 3)

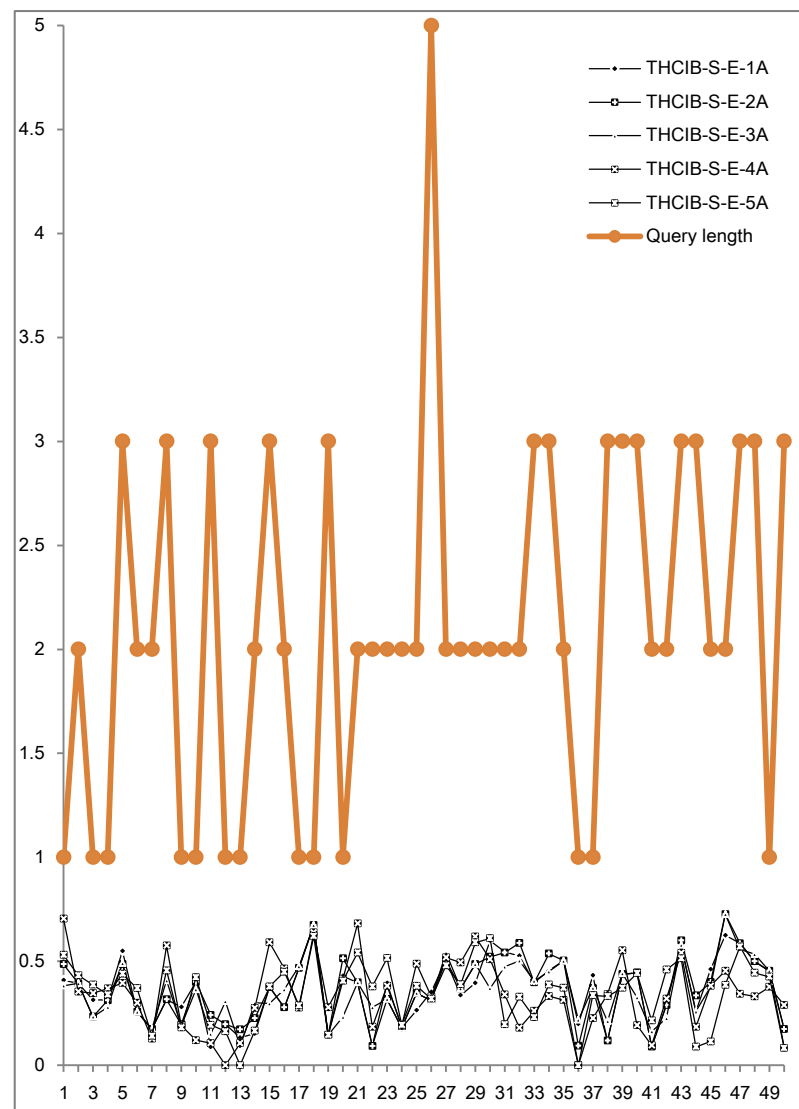
English Subtopic Mining runs

cut-off	run name	I-rec	D-nDCG	D#-nDCG
@10	THCIB-S-E-1A	0.3785	0.3384	0.3584
	THCIB-S-E-2A	0.3797	0.3499	0.3648
	THCIB-S-E-3A	0.3681	0.3383	0.3532
	THCIB-S-E-4A	0.3502	0.3323	0.3413
	THCIB-S-E-5A	0.3662	0.3215	0.3438
@20	THCIB-S-E-1A	0.5769	0.3274	0.4522
	THCIB-S-E-2A	0.5899	0.3406	0.4653
	THCIB-S-E-3A	0.5544	0.3251	0.4397
	THCIB-S-E-4A	0.477	0.2784	0.3777
	THCIB-S-E-5A	0.5395	0.304	0.4218
@30	THCIB-S-E-1A	0.693	0.3177	0.5054
	THCIB-S-E-2A	0.6743	0.3284	0.5014
	THCIB-S-E-3A	0.6486	0.3244	0.4865
	THCIB-S-E-4A	0.5855	0.2691	0.4273
	THCIB-S-E-5A	0.6339	0.2986	0.4662

- Performance in Chinese task is similar

Results and discussion – Per-topic analysis

- Best runs on the 50 queries
 - THCIB-S-E-1A 8
 - THCIB-S-E-2A 13
 - THCIB-S-E-3A 6
 - THCIB-S-E-4A 13
 - THCIB-S-E-5A 10
 - No run is consistently best, and each shows strength (further study is necessary)
- Query length
 - Our system is not sensitive to query length (Num. of words)
 - Other factors should be studied.



Outline

- The motivation
- System overview
- What make our system different?
- Evaluation
 - The submitted runs
 - Results and discussion
- **Conclusion and future work**

Conclusion and future work

- Conclusion
 - Incorporating **concepts** and **word senses** in subtopic mining and ranking brings marginal performance gain (NLP is positive to SM).
 - Subtopic ranking based on the automatically discovered intent is promising (**though** more work is required to improve intent quality).
- Future work
 - Deeper understanding the query: better subtopic extraction and intent discovery
 - Complexity issue: concept based indexing and retrieval
 - How about navigational and transactional query?

Acknowledgement

We thank Canon Inc. for supporting this research (No. TEMA2012).

We also thank the valuable comments from INTENT-2 organizer.



THANK YOU!

Q&A

We also welcome offline discussion by sending emails to
yqxia@tsinghua.edu.cn